CLAIMS

What is claimed is:

1. A method for performing a data-modifying operation in a file system that includes a plurality of servers that store replicas of data, one of the servers serving as a primary replica for one of the replicas of data and at least one other one of the servers serving as at least one secondary replica for the one replica of data, the method comprising:

sending data associated with the data-modifying operation to the primary replica and the at least one secondary replica based on a network topology; and

independently sending a data-modifying control signal that requests execution of the data-modifying operation using the data associated with the data-modifying operation to the primary replica and the at least one secondary replica.

2. The method of claim 1, wherein the sending data associated with the data-modifying operation includes:

pushing the data to one of the primary replica and the at least one secondary replica that is closest in the network topology to a sender of the data, the one of the primary replica and the at least one secondary replica serving as a closest replica.

3. The method of claim 2, wherein the sending data associated with the data-modifying operation further includes:

forwarding the data from the closest replica to one of the primary replica and the at least one secondary replica that is closest in the network topology to the closest replica.

4. The method of claim 3, wherein the sending data associated with the data-modifying operation further includes:

continuing to forward the data based on the network topology until all of the primary replica and the at least one secondary replica have received the data.

5. The method of claim 1, wherein the sending data associated with the data-modifying operation includes:

pipelining transmission of the data to the primary replica and the at least one secondary replica.

6. The method of claim 5, wherein the pipelining transmission of the data includes:

receiving the data at one of the primary replica and the at least one secondary replica, and

while receiving the data at the one of the primary replica and the at least one secondary replica, forwarding the data to another one of the primary replica and the at least one secondary replica.

7. The method of claim 1, wherein the independently sending a data-modifying control signal includes:

receiving the data-modifying control signal at the primary replica from a sender of the data-modifying control signal, and

forwarding the data-modifying control signal to the at least one secondary replica.

8. The method of claim 1, wherein the at least one secondary replica includes a plurality of secondary replicas; and

wherein the method further comprises:

assigning serial numbers to the secondary replicas that define an order in which the secondary replicas perform the data-modifying operation.

9. The method of claim 1, further comprising:

reporting that the data-modifying operation is successful when the data-modifying operation is successfully executed at the primary replica and the at least one secondary replica.

10. The method of claim 1, wherein the file system further includes a master; and

wherein the method further comprises:

granting a lease to one of the servers that stores the one replica of data, the one server thereafter being the primary replica.

11. The method of claim 10, wherein the lease has an initial timeout period.

12. The method of claim 11, wherein the initial timeout period is extendable.

13. The method of claim 10, further comprising:

receiving, by the master, a request for identification of the primary replica and the at least one secondary replica; and

sending, by the master, a reply that identifies locations of the primary replica and the at least one secondary replica.

14. A system for performing a data-modifying operation in a file network that includes a plurality of servers that store replicas of data, one of the servers serving as a primary replica for one of the replicas of data and other ones of the servers serving as secondary replicas for the one replica of data, the system comprising:

means for pushing data associated with the data-modifying operation to the primary replica and the secondary replicas based on a network topology; and

means for sending a data-modifying control signal to the primary replica and the at least one secondary replica independently of the pushing of the data, the data-modifying control signal requesting execution of the data-modifying operation using the data associated with the data-modifying operation.

15. A file system, comprising:

a primary replica server configured to store a replica of data; and

at least one secondary replica server configured to also store the replica of data, the primary replica server and the at least one secondary replica server in combination being configured to:

receive data associated with a data-modifying operation at one of the primary replica server and the at least one secondary replica server that is closest to a sender of the data,

forward the data to another one of the primary replica server and the at least one secondary replica server from the one of the primary replica server and the at least one secondary replica server that is closest to the sender of the data,

receive, at the primary replica server, a data-modifying control signal that requests execution of the data-modifying operation using the data associated with the data-modifying operation, and

forward the data-modifying control signal to the at least one secondary replica server from the primary replica server.

16. A method for performing a data-modifying operation in a file system that includes a plurality of servers that store replicas of data, one of the servers serving as a primary replica for one of the replicas of data and other ones of the servers serving as secondary replicas for the one replica of data, the method comprising:

receiving data associated with the data-modifying operation at one of the primary replica and the secondary replicas;

forwarding the data from the one of the primary replica and the secondary replicas to other ones of the primary replica and the secondary replicas;

receiving, at the primary replica, a data-modifying signal that requests execution of the data-modifying operation using the data associated with the data-modifying operation, the primary replica receiving the data-modifying signal independently of the data; and

forwarding the data-modifying signal to the secondary replicas.

17. A file system, comprising:

a plurality of servers configured to store replicas of data; and

a master connected to the servers and configured to:

receive a request for identification of the servers that store a replica of data,

determine whether one of the servers has a lease for the replica of data,

identify the one server as a primary replica when the one server has a lease for the replica of data,

identify other ones of the servers, as secondary replicas, that store the replica of data, and

send a reply that identifies locations of the primary replica and the secondary replicas.

18. The system of claim 17, wherein the master is further configured to grant a lease to one of the servers that stores the replica of data when none of the servers has a lease for the replica of data.

19. The system of claim 17, wherein the primary replica is configured to assign serial numbers to the secondary replicas that define an order in which the secondary replicas perform a data-modifying operation associated with the replica of data.

20. The system of claim 17, wherein the lease has an initial timeout period.

21. The system of claim 20, wherein the initial timeout period is extendable.

22. In a file system that includes a client connected to a master and a plurality of servers, the client comprising:

means for sending a request to the master for information regarding which of the servers store a replica of data;

means for receiving, from the master, identification of locations of at least first and second ones of the servers;

means for pushing data to at least one of the first and second servers by transmitting the data to a closest one of the first and second servers; and

means for transmitting, to the first sever, a control signal that instructs the first and second servers of an operation to perform upon the data.